Assignment 2

Artificial Datasets

Univariate Case

a) Generate 20 real number for the variable X from the uniform distribution U [0,1]
b) Construct the training set T = { $(x_1,y_1),(x_2,y_2),\ldots\ldots,(x_{20},y_{20})$} using the relation
   1. $Yi = \sin(2\,\pi\, x_i) + \epsilon_i$ where $\epsilon_i \sim N(0,0.25)$
c) In the similar way construct a testing set of size 50
   a. I,e. Test = { $(x'_1,y'_1),(x'_2,y'_2),\ldots\ldots,(x'_{50},y'_{50})$}
d) Estimate the regularized least squared polynomial regression model of order M= 1,2, 3, 9, using the training set T.
   i. For example for M=1 , we need to estimate
   ii. $F(x) = \beta_1 x + \beta_0$
   iii. For M = 2
   iv. $F(x) = \beta_2 x^2 + \beta_1 x + \beta_0$.

e) List the value of coefficients of estimated regularized least squared polynomial regression models for each case.
f) Obtain the prediction on testing set and compute the RMSE for regularized least squared polynomial regression models for order M =1,2,3 and 9.
g) Plot the estimate obtained by regularized least squared polynomial regression models for order M =1,2,3 and 9 for training set along with $y_{1,}\, y_{2,}$, $y_{20.}$. Also plot our actual mean estimate $E(Y/X) = \sin(2\,\pi\, x_i)$.
h) Plot the estimate obtained by regularized least squared polynomial regression models for order M =1,2,3 and 9 for testing set along with $y'_{1,}\, y'_{2,}$, $y'_{50.}$. Also plot the $\sin(2\,\pi\, x'_i)$.
i) Study the effect of regularization parameter λ on testing RMSE and flexibility of curve and list your observations.

Bivariate Case

a) Construct the training set T = { $(x_1,y_1),(x_2,y_2),\ldots\ldots,(x_3,y_{20})$} using the relation

$Y_i = \sin(2\pi(||x_i||) + \epsilon_i$ where $\epsilon_i \sim N(0,0.25)$ and $x_i = (x_i^1, x_i^2)$ where x_i^1, x_i^2 are from U[0,1]. In the similar way construct a testing set of size 50

a. I,e. Test = $\{(x'_1,y'_1),(x'_2,y'_2),\ldots\ldots,(x'_{50},y'_{50})\}$

b) Obtain the prediction on testing set and compute the RMSE for regularized least squared polynomial regression models for order M =1,2 and 5 . Also plot the estimated function and target function for the training set and testing set.

Real-world Datasets

a. Consider the motorcycle dataset. Estimate the Regularized Least Square regression models using the n sigmoidal basis functions. A variant of sigmoidal basis function can be obtained using

$$\sigma(a,b,x) = a^T x + b\,, a \in R^n, b \in R \; for \; x \in R^n.$$

I. Plot the estimated function and obtain the training RMSE error for n = 2, 5 , 10. What happens when you increase the number of basis functions.

II. For n =10, find the minimum mean and standard deviations of RMSE, NMSE and R2 using leave-one out method by tunning the parameter λ.

b. Consider the Boston Housing dataset. Use the ten-fold cross validation for obtaining the prediction of house price using the regularized least square RBF kernel regression model. By tunning the parameter λ and kernel parameter σ, obtain the minimum of mean of RMSE, NMSE, R^2, MAE, training times (in seconds) along with their standard deviation across different folds.